Minnesota Historical Society

Hubert H. Humphrey Digitization Project
Textual Digitization Workflow

The two main objectives of this workflow are to scan the entire series of Humphrey's speech texts and provide public access to those files by embedding the resulting PDFA-1b files as digital archival objects in an EAD inventory.

The workflow includes:

- Scan folders into PDF files using a Fujitsu fi-6230 sheet feeder for batch scanning and an Epson Expression 10000 XL flatbed scanner for delicate or differently-sized materials.
- Scan pages in 8-bit grayscale at a resolution of 300 dpi to maximize quality while minimizing file size.
- Produce thumbnail JPG images from the first scan of each file for an iconographic link in the inventory using both Microsoft Office Picture Manager and Adobe Bridge CS5.1.
- Save JPG files to an "images" subfolder inside a folder with the same filename as the EAD identifier.
- Run OCR program in Adobe Acrobat Pro software on PDF files to create an embedded layer of recognized text within the original file. Set OCR options to include Primary OCR language, Searchable PDF Output Style and No Downsampling.
- Save PDF files as compressed files. Set options to include Adobe Acrobat 7.0 and later compatibility, grayscale images, Bicubic Downsampling and retain existing compression.
- Save PDF files to a "pdfa" subfolder inside a folder with the same filename as the EAD identifier.
- Test quality of PDF scans by previewing files for errors, missed pages, and other occurrences. Correct any quality control issues.
- Add minimal metadata (Title, Author, Keyword and Copyright Information).
- Use Adobe Acrobat preflight function to output PDF/A-1b files.
- Extract EAD metadata from the XML finding aid, use daogrp template to add encoding for the PDFA files, and embed digital archival objects back into the XML inventory.